

Trace Norm Regularized Tensor Classification and Its Online Learning Approaches

Ziqiang Shi, Tieran Zheng, and Jiqing Han

September 8, 2011

Abstract

In this paper we propose an algorithm to classify tensor data. Our methodology is built on recent studies about matrix classification with the trace norm constrained weight matrix and the tensor trace norm. Similar to matrix classification, the tensor classification is formulated as a convex optimization problem which can be solved by using the off-the-shelf accelerated proximal gradient (APG) method. However, there are no analytic solutions as the matrix case for the updating of the weight tensors via the proximal gradient. To tackle this problem, the Douglas-Rachford splitting technique and the alternating direction method of multipliers (ADM) used in tensor completion are adapted to update the weight tensors. Further more, due to the demand of real applications, we also propose its online learning approaches. Experiments demonstrate the efficiency of the methods.

1 Introduction

Tensor or multi-way data analysis have many applications in the field of psychometrics, econometrics, image processing, signal processing, neuroscience, and data mining [1]. Tensors are higher-order equivalent of vectors and matrices. In this paper, we consider the classification of tensors, which is a generalization of the matrices classification problem proposed by Tomioka and Aihara in [2]. The tensor classification model is formulated as:

$$f(\mathcal{X}; \mathcal{W}, b) = \langle \mathcal{W}, \mathcal{X} \rangle + b \quad (1)$$

where $\mathcal{W}, \mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ are N -way tensors, \mathcal{X} is the input tensor for which we would like to predict its class label y ; \mathcal{W} is called the *weight tensor* and $b \in \mathbb{R}$ is the *bias*. Thus we need to infer the *weight tensor* and *bias* from the training samples $\{\mathcal{X}_i, y_i\}_{i=1}^s$. This formulation makes the work in [2] as a special case that the tensors evolved have an order of $N = 2$.

In the work of matrix classification, Tomioka and Aihara use a norm regularized scheme based on trace norm of the weight matrix [2]. Recently this trace norm regularization scheme has been studied in various contexts, namely, multi-task learning [3], matrix completion [4,5], and robust principle component analysis [6]. In this paper, similarly to matrix classification, a trace norm for tensors may be introduced to control the complexity of the weight tensor and the deviation of the empirical statistics from the predictions together. Recently, Liu et al. [7] proposed a definition for the tensor trace norm:

$$\|\mathcal{X}\|_* := \frac{1}{N} \sum_{i=1}^N \|X_{(i)}\|_* \quad (2)$$

where $X_{(i)}$ is the mode- i unfolding of \mathcal{X} , $\|X_{(i)}\|_*$ is the trace norm of the matrix $X_{(i)}$, i.e. the sum of the singular values of $X_{(i)}$, and if $N = 2$, this tensor norm is just the ordinary matrix trace norm. Now the weight tensor and bias learning problem becomes a convex optimization problem

$$\min_{\mathcal{W}, b} F_s(\mathcal{W}, b) = f_s(\mathcal{W}, b) + \lambda \|\mathcal{W}\|_*, \quad (3)$$

where $f_s(\mathcal{W}, b) = \sum_{i=1}^s \ell(y_i, \langle \mathcal{W}, \mathcal{X}_i \rangle + b)$ is the empirical cost function induced by some convex smooth loss function $\ell(\cdot, \cdot)$, and λ is the regularization parameter. The subscript of $f_s(\mathcal{W}, b)$ indicates the number of training samples or time of training procedure which is apparent from context.

For such convex optimization problem, Toh and Yun [8], Ji and Ye [9], and Liu et al. [10] independently proposed similar algorithms in the context of matrix related problems via using accelerated proximal gradient (APG) based methods. In this paper, we adapted the APG based algorithm to this tensor convex optimization problem. Unfortunately, unlike the Theorem 3.1 in [9] for matrix case, there is no closed analytic solution of the weight updating rules in the APG algorithm for the tensor case due to the dependency among multiple constraints. In order to solve the weight updating problem, the Douglas-Rachford splitting technique and the alternating direction method of multipliers [15,16], which have been successfully used in tensor completion tasks [7,11], are employed.

Furthermore, in order to cope with the situations that huge size training set for the data cannot be loaded into the memory simultaneously or the training data appear in sequence (for example video processing), we propose the online implementations of the above algorithms.

2 Notations

We adopt the nomenclature used by Kolda and Bader on tensor decompositions and applications [1]. The *order* N of a tensor is the number of dimensions, also known as ways or modes. Matrices (tensor of order two) are denoted by upper case letters, e.g. X , and lower case letters for the elements, e.g. x_{ij} . Higher-order tensors (order three or higher) are denoted by Euler script letters, e.g. \mathcal{X} , and element (i_1, i_2, \dots, i_N) of a N -order tensor \mathcal{X} is denoted by $x_{i_1 i_2 \dots i_N}$. *Fibers* are the higher-order analogue of matrix rows and columns. A fiber is defined by fixing every index but one. The mode- n fibers are all vectors $x_{i_1 \dots i_{n-1} : i_{n+1} \dots i_N}$ that obtained by fixing the values of $\{i_1, i_2, \dots, i_N\} \setminus i_n$. The mode- n *unfolding*, also known as *matricization*, of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$ is denoted by $X_{(n)}$ and arranges the mode- n fibers to be the columns of the resulting matrix. The unfolding operator is denoted as $\text{unfold}(\cdot)$. The opposite operation is $\text{refold}(\cdot)$, denotes the refolding of the matrix into a tensor. The tensor element (i_1, i_2, \dots, i_N) is mapped to the matrix element (i_n, j) , where

$$j = 1 + \sum_{\substack{k=1 \\ k \neq n}}^N (i_k - 1) J_k \quad \text{with} \quad J_k = \prod_{\substack{m=1 \\ m \neq n}}^{k-1} I_m$$

Therefore, $X_{(n)} \in \mathbb{R}^{I_n \times I_1 \dots I_{n-1} I_{n+1} \dots I_N}$. The n -*rank* of a N -dimensional tensor \mathcal{X} , denoted as $\text{rank}_n(\mathcal{X})$ is the column rank of $X_{(n)}$, i.e. the dimension of the vector space spanned by the mode- n fibers. The inner product of two same-size tensors $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$ is defined as

$$\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \dots \sum_{i_N=1}^{I_N} x_{i_1 i_2 \dots i_N} y_{i_1 i_2 \dots i_N}.$$

The corresponding norm is $\|\mathcal{X}\|_F = \sqrt{\langle \mathcal{X}, \mathcal{X} \rangle}$, which is often called the Frobenius norm.

3 Accelerated Proximal Gradient Method

It is known [8] that the gradient step

$$\mathcal{W}_k = \mathcal{W}_{k-1} - \frac{1}{t_k} \nabla_{\mathcal{W}} f_s(\mathcal{W}_{k-1}, b) \quad (4)$$

for solving the following smooth problem with fixed bias b

$$\min_{\mathcal{W}} f_s(\mathcal{W}, b) \quad (5)$$

without trace norm regularization can be formulated equivalently as a proximal regularization of the linearized function $f_s(\mathcal{W}, b)$ at \mathcal{W}_{k-1} as

$$\mathcal{W}_k = \operatorname{argmin}_{\mathcal{W}} P_{t_k}(\mathcal{W}, \mathcal{W}_{k-1}), \quad (6)$$

where

$$P_{t_k}(\mathcal{W}, \mathcal{W}_{k-1}) = f_s(\mathcal{W}_{k-1}, b) + \langle \mathcal{W} - \mathcal{W}_{k-1}, \nabla_{\mathcal{W}} f_s(\mathcal{W}_{k-1}, b) \rangle + \frac{t_k}{2} \|\mathcal{W} - \mathcal{W}_{k-1}\|_F^2 \quad (7)$$

and $\nabla_{\mathcal{W}} f_s(\cdot, b)$ is the gradient of $f_s(\cdot, b)$ with respect to \mathcal{W} .

Based on this equivalence relation, Toh and Yun [8], Ji and Ye [9], and Liu et al. [10] proposed to solve the optimization problem in Eq. (3) by the following iterative step:

$$\mathcal{W}_k = \operatorname{argmin}_{\mathcal{W}} Q_{t_k}(\mathcal{W}, \mathcal{W}_{k-1}) \triangleq P_{t_k}(\mathcal{W}, \mathcal{W}_{k-1}) + \lambda \|\mathcal{W}\|_* \quad (8)$$

or equivalently

$$\mathcal{W}_k = \operatorname{argmin}_{\mathcal{W}} \left\{ \frac{t_k}{2} \|\mathcal{W} - (\mathcal{W}_{k-1} - \frac{1}{t_k} \nabla_{\mathcal{W}} f_s(\mathcal{W}_{k-1}, b))\|_F^2 + \lambda \|\mathcal{W}\|_* \right\}. \quad (9)$$

Unfortunately, when the order of the tensor evolved in the problem is three or higher, there is no closed analytic solution to the above problem due to the tensor norm. This is contrast to the matrix case, where the Eq. (9) can be solved by singular value decomposition (SVD) and soft “shrinkage” like the theorem 3.1 in [9]. However, the Douglas-Rachford splitting technique and the alternating direction method of multipliers can be used to solve Eq. (9) for higher tensors. These methods will be described in the next section. Now, we assume that the Eq. (8) or Eq. (9) can be properly solved.

In general APG methods, the Lipschitz constant for $\nabla_{\mathcal{W}} f_s(\cdot, b)$ is unknown, so it is need to estimate the appropriate step size t_k to guarantee the convergence rate [8,9,10]. In this work, the standard squared loss function is used in Eq. (3). With this loss function, we can explicitly compute the Lipschitz constant in Lemma 3.1. Thus the step size estimation can be omitted in our tensor classification problems.

Lemma 3.1. $\nabla_{\mathcal{W}} f_s(\cdot, b)$ is Lipschitz continuous with constant $L = 2 \prod_{m=1}^N I_m \sum_{i=1}^s \|\mathcal{X}_i\|_F^2$, i.e.,

$$\|\nabla_{\mathcal{W}} f_s(\mathcal{U}, b) - \nabla_{\mathcal{W}} f_s(\mathcal{V}, b)\|_F \leq L \|\mathcal{U} - \mathcal{V}\|_F, \forall \mathcal{U}, \mathcal{V} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}, \quad (10)$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Proof. With the standard squared loss, the gradient of $f_s(\mathcal{W}, b)$ with respect to \mathcal{W} is

$$\nabla_{\mathcal{W}} f_s(\mathcal{W}, b) = -2 \sum_{i=1}^s (y_i - \langle \mathcal{W}, \mathcal{X}_i \rangle - b) \mathcal{X}_i, \quad (11)$$

Applying Eq. (11) with \mathcal{U}, \mathcal{V} to the right of Eq. (10), we obtain

$$\begin{aligned} & \|\nabla_{\mathcal{W}} f_s(\mathcal{U}, b) - \nabla_{\mathcal{W}} f_s(\mathcal{V}, b)\|_F \\ &= \left\| -2 \sum_{i=1}^s (y_i - \langle \mathcal{U}, \mathcal{X}_i \rangle - b) \mathcal{X}_i + 2 \sum_{i=1}^s (y_i - \langle \mathcal{V}, \mathcal{X}_i \rangle - b) \mathcal{X}_i \right\|_F \\ &= 2 \left\| \sum_{i=1}^s (\langle \mathcal{U}, \mathcal{X}_i \rangle - \langle \mathcal{V}, \mathcal{X}_i \rangle) \mathcal{X}_i \right\|_F \\ &\leq 2 \sum_{i=1}^s |\langle \mathcal{U} - \mathcal{V}, \mathcal{X}_i \rangle| \|\mathcal{X}_i\|_F \\ &\leq 2 \prod_{m=1}^N I_m \sum_{i=1}^s \|\mathcal{U} - \mathcal{V}\|_F \|\mathcal{X}_i\|_F^2 \\ &= (2 \prod_{m=1}^N I_m \sum_{i=1}^s \|\mathcal{X}_i\|_F^2) \|\mathcal{U} - \mathcal{V}\|_F, \end{aligned}$$

where in the last inequality, the easily verified fact that $\langle \mathcal{A}, \mathcal{B} \rangle \leq \|\mathcal{A}\|_1 \|\mathcal{B}\|_1 \leq \prod_{m=1}^N I_m \|\mathcal{A}\|_F \|\mathcal{B}\|_F$ for $\forall \mathcal{A}, \mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$ is used. Here $\|\cdot\|_1$ denotes the ℓ_1 norm which is the sum of the absolute values of the tensor elements.

Thus the lemma is proved, that is to say $\nabla_{\mathcal{W}} f_s(\cdot, b)$ is Lipschitz continuous with constant $L = 2 \prod_{m=1}^N I_m \sum_{i=1}^s \|\mathcal{X}_i\|_F^2$. \square

Based on the the work of Nesterov [13,14], Toh and Yun [8], Ji and Ye [9], and Liu et al. [10] showed that introduce a search point sequence $\mathcal{Z}_k = \mathcal{W}_k + \frac{t_{k-1}-1}{t_k}(\mathcal{W}_k - \mathcal{W}_{k-1})$ for a sequence t_k satisfying $t_{k+1}^2 - t_{k+1} \leq t_k^2$ results in a convergence rate of $O(\frac{1}{k^2})$. Based on their results, we adapted the APG algorithm to the tensor classification case and summarized in Algorithm 1. In this algorithm, the step of line 2 is not explicit. In the next section we will introduce some methods to solve this problem.

When the weight tensor is obtained, the bias b can be derived by solving the following problem with fixed weight tensor

$$b_k = \underset{b}{\operatorname{argmin}} \left\{ \sum_{i=1}^s (y_i - \langle \mathcal{W}_k, \mathcal{X}_i \rangle - b)^2 + \lambda \|\mathcal{W}_k\|_* \right\}, \quad (12)$$

which results in the bias updating rule

$$b_k = \frac{1}{s} \sum_{i=1}^s (y_i - \langle \mathcal{W}_k, \mathcal{X}_i \rangle). \quad (13)$$

4 Minimization via Gandy's Algorithms

Apparently that the problem of Eq. (9) or line 2 in Algorithm 1 fulfils the recently proposed tensor completion formulation [7,11]. For tensor completion, Gandy proposed two algorithms

Algorithm 1 Weight Tensor Learning via APG

Input $(\mathcal{X}_i, y_i), i = 1, \dots, s$.

Initialization $\mathcal{W}_0 = \mathcal{Z}_1 \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}, \alpha_1 = 1, L = 2 \prod_{m=1}^N I_m \sum_{i=1}^s \|\mathcal{X}_i\|_F^2, \lambda, k = 1$.

1: **while** not converged **do**

2: $\mathcal{W}_k = \underset{\mathcal{W}}{\operatorname{argmin}} \{ \frac{L}{2} \|\mathcal{W} - (\mathcal{Z}_k - \frac{1}{L} \nabla_{\mathcal{W}} f_s(\mathcal{Z}_k, b))\|_F^2 + \lambda \|\mathcal{W}\|_* \}$.

3: $\alpha_{k+1} = \frac{1 + \sqrt{1 + 4\alpha_k^2}}{2}$.

4: $\mathcal{Z}_{k+1} = \mathcal{W}_k + \frac{\alpha_k - 1}{\alpha_{k+1}} (\mathcal{W}_k - \mathcal{W}_{k-1})$.

5: $k \leftarrow k + 1$.

6: **end while**

Output: $\mathcal{W} \leftarrow \mathcal{W}_k$.

based on Douglas-Rachford splitting technique and the alternating direction method of multipliers (ADM) respectively. In this work, we adapt these two methods to solve the problem (9).

Douglas-Rachford splitting technique based method: The Douglas-Rachford splitting technique has a long history [15,16]. It addresses the minimization of the sum of two functions $(f + g)(x)$, where f and g are lower semicontinuous convex functions. The Douglas-Rachford splitting technique asserted that $\operatorname{prox}_{\lambda g}(\tilde{x})$ is a minimizer of $(f + g)(x)$, where \tilde{x} is the limit point of the following sequence:

$$x_{n+1} := x_n + t_n \{ \operatorname{prox}_{\lambda f} [2 \operatorname{prox}_{\lambda g}(x_n) - x_n] - \operatorname{prox}_{\lambda g}(x_n) \}, \quad (14)$$

where $t_n \in [0, 2]$ satisfies $\sum_{n \geq 0} t_n(2 - t_n) = \infty$ and the proximal map $\operatorname{prox}_{\lambda g}(\cdot)$ is defined as [17,18]:

$$\operatorname{prox}_{\lambda f} : x \mapsto \underset{y}{\operatorname{argmin}} \{ f(y) + \frac{1}{2\lambda} \|x - y\|^2 \}. \quad (15)$$

We first formulate the problem in step 2 of Algorithm 1 into the unconstrained minimization of $(f + g)(x)$. Let $\mathfrak{F} := \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$, define a Hilbert space $\mathfrak{H}_0 := \underbrace{\mathfrak{F} \times \mathfrak{F} \times \dots \times \mathfrak{F}}_{N+1 \text{ terms}}$ with the inner

product $\langle \mathfrak{X}, \mathfrak{Y} \rangle_{\mathfrak{H}_0} := \frac{1}{N+1} \sum_{i=0}^N \langle \mathcal{X}_i, \mathcal{Y}_i \rangle$. Then the problem can be rephrased as:

$$\underset{\mathfrak{W} \in \mathfrak{H}_0}{\operatorname{minimize}} \quad f(\mathfrak{W}) + g(\mathfrak{W}), \quad (16)$$

where $\mathfrak{W} = (\mathcal{W}_0, \mathcal{W}_1, \dots, \mathcal{W}_N)$, $D = \{ \mathfrak{W} \in \mathfrak{H}_0 | \mathcal{W}_0 = \mathcal{W}_1 = \dots = \mathcal{W}_N \}$, and

$$f(\mathfrak{W}) = \frac{L}{2} \|\mathcal{W}_0 - \mathcal{P}\|_F^2 + \sum_{i=1}^N \frac{\lambda}{N} \|W_{i,(i)}\|_*, \quad (17)$$

$$g(\mathfrak{W}) = i_D(\mathfrak{W}) = \begin{cases} 0, & \text{if } \mathfrak{W} \in D \\ +\infty, & \text{otherwise} \end{cases} \quad (18)$$

where $\mathcal{P} = \mathcal{Z}_{k-1} - \frac{1}{L} \nabla_{\mathcal{W}} f_s(\mathcal{Z}_{k-1}, b)$. Then in order to apply the stand DR splitting technique, the proximal maps of $f(\mathfrak{W})$ and $g(\mathfrak{W})$ need to be identified.

The proximal map of $f(\mathfrak{W})$ is given by

$$\begin{aligned}
\text{prox}_{\gamma f} \mathfrak{W} &= \arg \min_{\mathfrak{W} \in \mathfrak{H}_0} \left\{ \frac{L}{2} \|\mathcal{W}_0 - \mathcal{P}\|_F^2 + \sum_{i=1}^N \frac{\lambda}{N} \|W_{i,(i)}\|_* + \frac{1}{2\gamma} \|\mathfrak{W} - \mathfrak{W}_0\|_{\mathfrak{H}_0}^2 \right\} \\
&= \arg \min_{\mathfrak{W} \in \mathfrak{H}_0} \left\{ \frac{L}{2} \|\mathcal{W}_0 - \mathcal{P}\|_F^2 + \sum_{i=1}^N \frac{\lambda}{N} \|W_{i,(i)}\|_* + \frac{1}{2(N+1)\gamma} \sum_{i=0}^N \|\mathcal{Y}_i - \mathcal{W}_i\|_F^2 \right\} \\
&= (\text{prox}_{(N+1)\gamma(\frac{L}{2}\|\mathcal{W}-\mathcal{P}\|_F^2)} \mathcal{W}_0, \text{prox}_{(N+1)\gamma(\frac{\lambda}{N}\|W_{1,(1)}\|_*)} \mathcal{W}_1, \dots, \text{prox}_{(N+1)\gamma(\frac{\lambda}{N}\|W_{N,(N)}\|_*)} \mathcal{W}_N)
\end{aligned}$$

For $\text{prox}_{(N+1)\gamma(\frac{L}{2}\|\mathcal{W}-\mathcal{P}\|_F^2)} \mathcal{W}_0$, we have

$$\arg \min_{\mathcal{W} \in \mathfrak{S}} \left\{ \frac{L}{2} \|\mathcal{W} - \mathcal{P}\|_F^2 + \frac{1}{2(N+1)\gamma} \|\mathcal{W} - \mathcal{Y}_0\|_F^2 \right\} = (\frac{L}{2} \mathcal{P} + \frac{1}{2(N+1)\gamma} \mathcal{Y}_0) / (\frac{L}{2} + \frac{1}{2(N+1)\gamma}). \quad (19)$$

For $\text{prox}_{(N+1)\gamma(\frac{\lambda}{N}\|W_{i,(i)}\|_*)} \mathcal{W}_i, i = 1, \dots, N$, by Theorem 3.1 in [9], we have

$$\arg \min_{\mathcal{W} \in \mathfrak{S}} \left\{ \frac{\lambda}{N} \|W_{i,(i)}\|_* + \frac{1}{2(N+1)\gamma} \|\mathcal{W} - \mathcal{Y}_i\|_F^2 \right\} = \text{refold}(U \mathcal{S}_{\frac{\lambda(N+1)\gamma}{N}}[S] V^T), \quad (20)$$

where USV^T is the SVD of $Y_{i,(i)}$, the $\text{refold}(\cdot)$ is referred to Section 2, and the $\mathcal{S}_\varepsilon[\cdot]$ is the soft-thresholding operator introduced in [19]:

$$\mathcal{S}_\varepsilon[x] \doteq \begin{cases} x - \varepsilon, & \text{if } x > \varepsilon, \\ x + \varepsilon, & \text{if } x < -\varepsilon, \\ 0, & \text{otherwise} \end{cases} \quad (21)$$

where $x \in \mathbb{R}$ and $\varepsilon > 0$. For vectors and matrices, this operator is extended by applying element-wise.

The proximal map of the indicator function $g(\mathfrak{W})$ is simply given by

$$\text{prox}_{\gamma g} \mathfrak{W} = (\widehat{\mathfrak{W}}, \dots, \widehat{\mathfrak{W}}),$$

where $\widehat{\mathfrak{W}} = \frac{1}{N+1} \sum_{i=1}^N \mathcal{W}_i$.

Now apply Eq. (14), we obtain the iteration rules for the original problem:

$$\mathcal{W}_0^{k+1} = \mathcal{W}_0^k + \arg \min_{\mathcal{W}} \left(\frac{L}{2} \|\mathcal{W} - \mathcal{P}\|_F^2 + \frac{1}{2(N+1)\gamma} \|\mathcal{W} - (2\widehat{\mathfrak{W}} - \mathcal{W}_0^k)\|_F^2 \right) - \widehat{\mathfrak{W}}, \quad (22)$$

$$\mathcal{W}_i^{k+1} = \mathcal{W}_i^k + \arg \min_{\mathcal{W}} \left(\frac{\lambda}{N} \|W_{i,(i)}\|_* + \frac{1}{2(N+1)\gamma} \|\mathcal{W} - (2\widehat{\mathfrak{W}} - \mathcal{W}_i^k)\|_F^2 \right) - \widehat{\mathfrak{W}}, i = 1, \dots, N. \quad (23)$$

The convergence is guaranteed by Theorem 4.1 in [11]. When it converges, the weight tensor is \mathfrak{W} .

ADM based method: The ADM based method goes back to last century [20]. The approach consists of iteratively updating the original variables and finally carrying out the update of the dual variables. Each update involves a single variable and is conditioned to the fixed value of the others. In order to use the ADM in tensor completion, Gandy introduced N new tensor-value variables that represents the N different mode- n unfoldings of the original tensor, then form the

augmented Lagrangian and update all the variables one at a time. Following Gandy's method, we introduce N new variable $\mathcal{Y}_i \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$ and rephrase line 2 of the Algorithm 1 as

$$\begin{aligned} \min_{\mathcal{W}, \mathcal{Y}_i} \quad & \frac{L}{2} \|\mathcal{W} - \mathcal{P}\|_F^2 + \frac{\lambda}{N} \sum_{i=1}^N \|\mathcal{Y}_{i,(i)}\|_* \\ \text{subject to} \quad & \mathcal{Y}_i = \mathcal{W} \quad \forall i \in \{1, \dots, N\}. \end{aligned} \quad (24)$$

Let $f(\mathcal{W}) = \frac{L}{2} \|\mathcal{W} - \mathcal{P}\|_F^2$, $g(\mathfrak{Y}) = \frac{\lambda}{N} \sum_{i=1}^N \|\mathcal{Y}_{i,(i)}\|_*$, where $\mathfrak{Y} = (\mathcal{Y}_1, \dots, \mathcal{Y}_N)^T$. Thus the constrain over \mathfrak{Y} and \mathcal{W} is $\mathfrak{Y} = (\mathcal{W}, \dots, \mathcal{W})$. Then the augmented Lagrangian of Eq. (24) becomes

$$\mathcal{L}_A(\mathcal{W}, \mathfrak{Y}, \mathfrak{U}) = \frac{L}{2} \|\mathcal{W} - \mathcal{P}\|_F^2 + \sum_{i=1}^N \left(\frac{\lambda}{N} \|\mathcal{Y}_{i,(i)}\|_* - \langle \mathcal{U}_i, \mathcal{W} - \mathcal{Y}_i \rangle + \frac{\beta}{2} \|\mathcal{W} - \mathcal{Y}_i\|_F^2 \right) \quad (25)$$

where the parameter β is any positive number and $\mathfrak{U} = (\mathcal{U}_1, \dots, \mathcal{U}_N)^T$ is the Lagrange multiplier. By minimization $\mathcal{L}_A(\mathcal{W}, \mathfrak{Y}, \mathfrak{U})$ with respect to each single variable and other variables fixed, we obtain the updating rules of all the variables $\mathfrak{Y}, \mathcal{W}, \mathfrak{U}$

$$\begin{cases} \mathcal{W}^{k+1} = (L\mathcal{P} + \beta \sum_{i=1}^N \mathcal{Y}_i + \sum_{i=1}^N \mathcal{U}_i) / (L + \beta N), \\ \mathcal{Y}_i^{k+1} = \text{refold}(US_{\frac{\lambda}{\beta N}}[S]V^T), i = 1, \dots, N, \\ \mathcal{U}_i^{k+1} = \mathcal{U}_i^k - \beta(\mathcal{W}^{k+1} - \mathcal{Y}_i^{k+1}), i = 1, \dots, N, \end{cases} \quad (26)$$

where USV^T is the SVD of $(W_{(j)}^{k+1} - \frac{1}{\beta} U_{j,(j)}^k)$.

Until now we have proposed two methods to solve the tensor classification problem. In the next section, we discuss the online implementation of the proposed learning processes.

5 Online Learning

The above proposed methods are iterative *batch* procedures, accessing the whole training set at each iteration in order to minimize a weighted sum of a cost function and the tensor trace norm. This kind of learning procedure cannot deal with huge size training set for the data probably cannot be loaded into memory simultaneously, furthermore it cannot be started until the training data are prepared, hence cannot effectively deal with the training data appear in sequence, such as audio and video processing.

To address these problems, we propose an *online* approach that processes the training samples, one at a time, or in mini-batches to learn the weight tensor and the bias for tensor classification. We transform the above algorithm to the online learning framework. The framework is described in Algorithm 2 in which we also include the bias updating steps.

Our procedure is summarized in Algorithm 2. The \otimes operator in step 6 of the algorithm denotes the Kronecker product which is similar to matrix Kronecker product. Given two tensors $\mathcal{A} \in \mathbb{R}^{I_1 \times \dots \times I_N}$ and $\mathcal{B} \in \mathbb{R}^{J_1 \times \dots \times J_N}$ with equal order N , $\mathcal{A} \otimes \mathcal{B}$ denotes the Kronecker product between \mathcal{A} and \mathcal{B} , results as a tensor in $\mathbb{R}^{I_1 J_1 \times \dots \times I_N J_N}$, defined by blocks of sizes $J_1 \times \dots \times J_N$ equal to $a_{i_1 \dots i_N} \mathcal{B}$. $\text{GridTr}(\mathcal{W}, \mathcal{B}_t)$ in step 13 denotes an operator with input $\mathcal{W} \in \mathbb{R}^{I_1 \times \dots \times I_N}$ and $\mathcal{B}_t \in \mathbb{R}^{I_1 J_1 \times \dots \times I_N J_N}$, result in $\mathbb{R}^{I_1 \times \dots \times I_N}$ with the (i_1, \dots, i_N) th element defined as the inner product between \mathcal{W} and the (i_1, \dots, i_N) th $\mathbb{R}^{I_1 \times \dots \times I_N}$ block of \mathcal{B}_t .

Algorithm 2 Online learning for tensor classification via APG

Initialization $\mathcal{W}_0 = 0 \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$, $b_0 \in \mathbb{R}$, λ .
1: $\mathcal{A}_0 \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N} \leftarrow 0$, $\mathcal{B}_0 \in \mathbb{R}^{I_1 I_2 \times \dots \times I_N I_N} \leftarrow 0$, $c_0 \in \mathbb{R} \leftarrow 0$, $\mathcal{D}_0 \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N} \leftarrow 0$, $L_0 = 0 \in \mathbb{R}$ (reset the “past” information).
2: **for** $t = 1$ **to** T **do**
3: Draw training sample (\mathcal{X}_t, y_t) from $p(\mathcal{X}, y)$.
4: // Line 5-9 update “past” information.
5: $\mathcal{A}_t \leftarrow \mathcal{A}_{t-1} + y_t \mathcal{X}_t$;
6: $\mathcal{B}_t \leftarrow \mathcal{B}_{t-1} + \mathcal{X}_t \otimes \mathcal{X}_t$;
7: $c_t \leftarrow c_{t-1} + y_t$;
8: $\mathcal{D}_t \leftarrow \mathcal{D}_{t-1} + \mathcal{X}_t$;
9: $L_t \leftarrow L_{t-1} + 2 \prod_{m=1}^N I_m \|\mathcal{X}_t\|_F^2$.
10: // Line 11-19 compute \mathcal{W}_t using the APG method, with \mathcal{W}_{t-1} as warm restart.
11: $\mathcal{W}_{0,t} = \mathcal{Z}_{1,t} = \mathcal{W}_{t-1} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$, $b_{0,t} = b_{t-1}$, $\alpha_1 = 1$, $k = 1$.
12: **while** not converged **do**
13: $\mathcal{W}_{k,t} = \underset{\mathcal{W}}{\operatorname{argmin}} \frac{L_t}{2} \|\mathcal{W} - (\mathcal{Z}_{k,t} + \frac{2}{L}(\mathcal{A}_t - \operatorname{GridTr}(\mathcal{Z}_{k,t}, \mathcal{B}_t) - b_{k-1,t} \mathcal{D}_t))\|_F^2 + \lambda \|\mathcal{W}\|_*$.
14: $\alpha_{k+1} = \frac{1 + \sqrt{1 + 4\alpha_k^2}}{2}$.
15: $\mathcal{Z}_{k+1,t} = \mathcal{W}_{k,t} + \frac{\alpha_k - 1}{\alpha_{k+1}}(\mathcal{W}_{k,t} - \mathcal{W}_{k-1,t})$.
16: $b_{k,t} = \frac{1}{t}(c_t - \langle \mathcal{W}_{k,t}, \mathcal{D}_t \rangle)$
17: $k \leftarrow k + 1$.
18: **end while**
19: $\mathcal{W}_t \leftarrow \mathcal{W}_{k,t}$, $b_t \leftarrow b_{k,t}$.
20: **end for**
Output: $\mathcal{W} \leftarrow \mathcal{W}_T$, $b \leftarrow b_T$.

Assuming the training set composed of i.i.d. samples of a distribution $p(\mathcal{X}, y)$, its inner loop draws one training sample (\mathcal{X}_t, y_t) at a time. This sample is first used to update the “past” information \mathcal{A}_{t-1} , \mathcal{B}_{t-1} , c_{t-1} , \mathcal{D}_{t-1} , L_{t-1} . Then the Algorithm 1 is applied to update the weight matrix with the warm start \mathcal{W}_{t-1} obtained at the previous iteration. Since $F_t(\mathcal{W}, b_{t-1})$ is relative close to $F_{t-1}(\mathcal{W}, b_{t-1})$ for large values of t , so are \mathcal{W}_t and \mathcal{W}_{t-1} , under suitable assumptions, which makes it efficient to use \mathcal{W}_{t-1} as warm restart for computing \mathcal{W}_t .

For the stopping criteria of the inside iterations, we take the following relative error conditions:

$$\|\mathcal{W}_{k+1,t} - \mathcal{W}_{k,t}\|_F / (\|\mathcal{W}_{k,t}\|_F + 1) < \varepsilon_1 \text{ and } |b_{k+1,t} - b_{k,t}| / (|b_{k,t}| + 1) < \varepsilon_2. \quad (27)$$

In some conditions, use the classical heuristic in gradient descent algorithm, we may also improve the convergence speed of our algorithm by drawing $\mu > 1$ training samples at each iteration instead of a single one. Let us denote by $(\mathcal{X}_{t,1}, y_{t,1}), \dots, (\mathcal{X}_{t,\mu}, y_{t,\mu})$ the samples drawn at iteration t . We can now replace lines 5 and 9 of Algorithm 2 by

$$\begin{aligned} \mathcal{A}_t &\leftarrow \mathcal{A}_{t-1} + \sum_{i=1}^{\mu} y_{t,i} \mathcal{X}_{t,i}, \quad \mathcal{B}_t \leftarrow \mathcal{B}_{t-1} + \sum_{i=1}^{\mu} \mathcal{X}_{t,i} \otimes \mathcal{X}_{t,i}, \quad c_t \leftarrow c_{t-1} + \sum_{i=1}^{\mu} y_{t,i}, \\ \mathcal{D}_t &\leftarrow \mathcal{D}_{t-1} + \sum_{i=1}^{\mu} \mathcal{X}_{t,i}, \text{ and } L_t \leftarrow L_{t-1} + \sum_{i=1}^{\mu} 2 \prod_{m=1}^N I_m \|\mathcal{X}_{t,i}\|_F^2. \end{aligned} \quad (28)$$

But in real applications, this online with mini-batch update method may not improve the convergence speed on the whole since the batch past information computation (Eq. (28)) would occupy

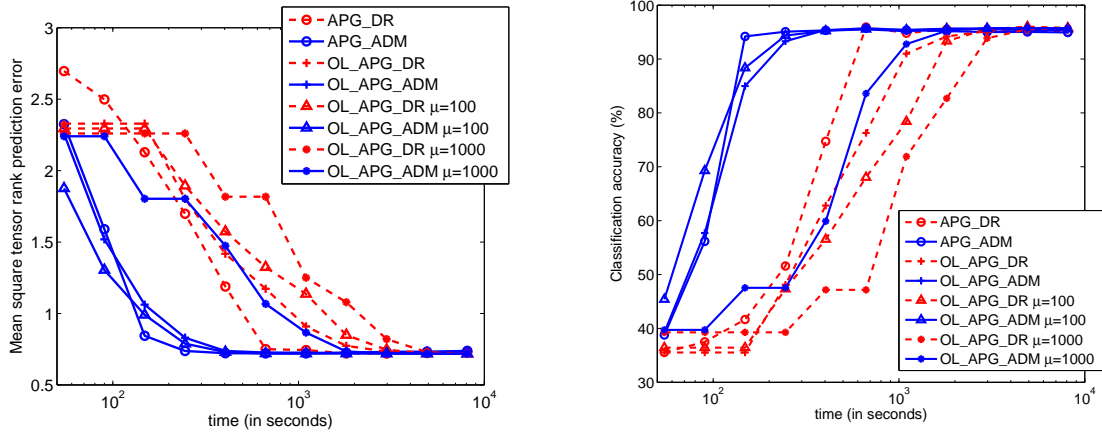
much of the time. The updating of \mathcal{B}_t needs to do Kronecher product which spend much of the computing resource. If the computation cost of Eq. (28) can be ignored or largely decreased, for example by parallel computing, this mini-batch method would increase the convergence speed by a factor of μ .

6 Experimental Validation

In this section, we conduct experiments to demonstrate the characteristics of the proposed methods for tensor classification problem. Six algorithms are compared: the batch learning algorithm with APG using DR methods (APG_DR); the online learning algorithm with APG using DR (OL_APG_DR); the batch learning algorithm with APG using ADM method (APG_ADM); the online learning algorithm with APG using ADM (OL_APG_ADM); OL_APG_DR with update Eq. (28) (OL_APG_DR_miniBatch); OL_APG_ADM with update Eq. (28) (OL_APG_ADM_miniBatch). All algorithms are run in Matlab on a PC with an Intel 2.53GHz dual-core CPU and 3.25GB memory.

For our experiments, we use randomly generated 2.4×10^5 3-order $10 \times 10 \times 10$ tensors, which are composed of varied ranks (note that here the rank is not the n -rank mentioned above, here the rank concept related to CANDECOMP/PARAFAC decomposition, refer [1] for exact definition); 2×10^5 of these are kept for training, and the rest for testing. The goal is to classify the tensors according to their ranks. Hence we have made the tensor rank identification problem into a novel classification or regression formulation. We generate the rank- r tensor as a sum of r rank one tensors, where each rank one tensor is a outer product of 3 vectors whose elements are drawn i.i.d from the standard uniform distribution on the open interval $(0, 1)$. For all the algorithm, the parameters in the stopping criteria (27) are $\varepsilon_1 = 10^{-10}$ and $\varepsilon_2 = 10^{-10}$. The regularization constant λ is anchored by the large explicit fixed step size L and the tensors involved, which means that in practice the parameter λ should be set adaptably with the step size L in the online process. But due to this variation of λ , the comparisons between the algorithms would not bring into effect. Hence in this work we use $\lambda = 1$ throughout. Considering a balance between convergence speed and accuracy, we set $\beta = 10^7, \gamma = 10^{-7}$ in this work.

Figure 1 compares all the algorithms proposed in this work. The batch algorithm use a training set of 2×10^3 training samples, while the online algorithm draws samples from the entire training set. We use a logarithmic scale for the computation time. Figure 1(a) shows the mean square tensor rank prediction errors as functions of time. It can be seen generally that all methods converge. In all these methods, ADM based methods converge faster than DR based methods. The batch learning methods converge faster than corresponding online learning methods with or without mini-batch past information updating. It can also be seen that when the size of the mini-batch used in online method increase, the speed of convergence will decrease, and the reason for this has been explained in the last paragraph of Section 5. After all the methods converge, they result in almost equal performance. Figure 1(b) shows the classification rates with tensor rank estimation error tolerances $\eta = 1$. Here the rank estimation error tolerance means that if the distance between the estimation rank value and the real rank value is less than η , then the tensor classification would be right. The convergence of the classification accuracies are corresponding to the convergence of the mean square tensor rank prediction errors. With an error tolerance $\eta = 1$, the methods result in a classification rate of 95.9%.



(a) Mean square rank prediction error as function of time. (b) Tensor classification accuracy with $\eta = 1$ as function of time.

Figure 1: Comparison between various learning methods and results are reported as functions of learning time on a logarithmic scale.

7 Conclusions

In this paper, we have proposed methods to solve tensor classification problem with a tensor trace norm regularization. We successfully employed APG method to learn parameters, during which DR and ADM are used to update weight tensor. We also give out online learning implementation for all proposed methods. In addition, for standard squared loss function, we derive the explicit form of the Lipschitz constant, which saves the computation burden in searching step size. Our empirical study on tensor classification according to tensor rank demonstrates the merits of the proposed algorithms. This is, to our knowledge, the first work on tensor norm constrained tensor classification. Some future work are worth considering, such as that the alternating between minimization with respect to weight tensor and bias may results in fluctuation of target value, thus optimization algorithm that minimization jointly on weight tensor and bias are required; for multi-classification problems with more classes, some hierarchy methods may be introduced to improve the classification accuracy.

References

- [1] Kolda, T.G. & Bader, B.W. (2009) Tensor decompositions and applications. *SIAM Review* **51**(3):455-500.
- [2] Tomioka, R. & Aihara, K. (2007) Classifying matrices with a spectral regularization. *24th International Conference on Machine Learning*, pp. 895-902.
- [3] Argyriou, A. & Evgeniou, T. & Pontil, M. (2008) Convex multi-task feature learning. *Machine Learning* **73**(3):243-272.
- [4] Srebro, N. & Rennie, J. D. M. & Jaakkola, T. S. (2005) Maximum-margin matrix factorization. *Proceedings of Advances in Neural Information Processing Systems*, pp. 1329-1336
- [5] Candes, E. J. & Recht, B. (2008) Exact matrix completion via convex optimization. *Technical Report, UCLA Computational and Applied Math*.

- [6] Wright, J. & Ganesh, A. & Rao, S. & Peng, Y. & Ma, Y. (2009) Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. *Proceedings of Advances in Neural Information Processing Systems*.
- [7] Liu, J., Musialski, P., Wonka, P. & Ye, J. (2009) Tensor completion for estimating missing values in visual data. *IEEE 12th International Conference on Computer Vision*, pp. 2114-2121.
- [8] Toh, K. & Yun, S. (2010) An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *Pacific J. Optim* **6**:615-640.
- [9] Ji, S. & Ye, J. (2009) An accelerated gradient method for trace norm minimization. *26th International Conference on Machine Learning*, pp. 457-464.
- [10] Liu, Y.J., Sun, D. & Toh, K.C. (2009) An implementable proximal point algorithmic framework for nuclear norm minimization. *Mathematical Programming*, pp. 1-38.
- [11] Gandy, S., Recht, B., & Yamada, I. (2011) Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems* **27**(2).
- [12] Bertsekas, D.P. (1999) *Nonlinear programming*. Athena Scientific Belmont, MA.
- [13] Nesterov, Y. (1983) A method of solving a convex programming problem with convergence rate $O(\frac{1}{k_2})$. *Soviet Mathematics Doklady*. **27**(2):372-376.
- [14] Nesterov, Y. (2005) Smooth minimization of non-smooth functions. *Mathematical Programming*. **103**(1):127-152.
- [15] Douglas, J. & Rachford, H. (1956) On the numerical solution of heat conduction problems in two and three space variables. *Trans. of the American Mathematical Society* **82**:421-439.
- [16] Combettes, P. L. & Pesquet, J. C. (2007) A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE J. Sel. Top. Signal Process* **1**(4):564-574.
- [17] Moreau, J.J. (1962) Fonctions convexes duales et points proximaux dans un espace hilbertien. *C.R.Acad.Sci. Paris Ser. A Math* **244**:2897-2899.
- [18] Combettes, P. L. & Wajs, V.R. (2005) Signal recovery by proximal forward-backward splitting. *SIAM Multiscale Model. Simul.* **4**:1168-1200.
- [19] Lin, Z., Chen, M., Wu, L. & Ma, Y. (2009) The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. preprint.
- [20] Gabay, D. & Mercier, B. (1976) A dual algorithm for the solution of nonlinear variational problems via finite-element approximations. *Comp. Math. Appl.* **2**:17-40.